图书情报工作
第 62 卷 第 13 期 2018 年 7 月

基于 Doc2Vec 的专利文件相似度检测方法的对比研究*

曹祺[1] 赵伟[1] 张英杰[2] 赵树君[3] 陈亮[4]
[1,2,4] 中国科学技术信息研究所 北京 100038 [3] 武汉大学 武汉 430072

摘要：[目的/意义]专利相似度检测(Similarity Measurement)可从宏观上辅助制定国家创新战略规划,发现国内外的热点及应对其他国家的专利流氓,从微观上为专利发明人、专利审查员、专利权人提供辅助支撑。[方法/过程]提出基于深度学习的 Doc2Vec 专利相似度分析方法,基于未进行清洗的专利语料库,采用深度学习的 Doc2Vec 模型,随机挑选了专利,研究了专利相似度检测问题,并和传统的相似度检测模型进行对比研究。[结果/结论]实验结果表明,基于深度学习的 Doc2Vec 模型和 TF-IDF 模型对于处理不做数据清洗的专利语料的结果有相近性,该方法对分析人员的专利领域知识要求较低,不需要对专利数据进行基于专利领域知识的数据清洗,同时可为专利侵权、专利查新提供新的智能工具支撑,降低研究门槛和工作量,提升研究效率。

关键词：专利 相似度 深度学习 Doc2Vec
分类号：G354
DOI：10.13266/j.issn.0252-3116.2018.13.010

1 引言

专利相似度检测在宏观和微观上有很重要的研究意义:从宏观上来说,一方面分析专利相似度,能辅助制定国家创新战略规划,发现国内外的热点。这方面,S. Mukherjee 等[1]利用 1945 年至 2013 年的 28 426 345 篇论文和 1950 年至 2010 年的 5 382 833 篇美国专利对比分析,提出了研究热点的定义并对比了不同阶段的研究热点及趋势演化。另一方面,发达国家经常利用知识产权对发展中国家进行不正当起诉,S. Padmanabhan 等[2]分析了发达国家利用 HPV 相关专利在印度进行起诉的相关流程,我国作为发展中国家,也需要有相关的应对措施,王曰芬等[3]研究了目前面向专利预警的专利文献相似度研究现状。从微观上来说,专利的发明人往往是某一细分学科的技术专家,撰写发明专利的目的是利用新技术对于所在细分学科的新产品方法进行保护,因此很难撰写交叉学科的发明专利。另一方面,专利审查员往往也是某一细分学科的技术专家,专利审查员进行专利审查时如果遇到交叉学科的专利也往往很难快速审查,需要花费大量时间对不同学科的知识进行学习。当前专利文本相似度检测的主流方法仍然是基于专利审查员的人工定义词库后进行词频检测,需要较高的专业背景或花费大量的时间精力去了解相关专利领域的知识。而最近几年,深度学习在文本相似度检测等领域得到了广泛的应用,专利也是一种特别的文本,也适用于采用深度学习的方法开展研究。本文采用深度学习的新技术,以美国专利为例,研究了专利相似度检测问题。

2 研究现状

2.1 专利文件结构及相似度分析流程

本文的研究对象是专利文件。以美国专利文件为例,美国专利的结构主要分为专利名、摘要,权利要求书、说明书、引文,见图 1[4]。

基于这样的结构树,目前专利相似度分析研究流程见图 2[5]。

通常做法中,不同的专利相似度分析流程的差别在于数据源选择策略、数据源分析算法等。这类相似度分析需要专利领域的技术专家具备多种专利的领域知识。在数据源的选择上,主要基于专利数据库和其他数据库跨库对比和专利数据库自身分析两种方法,

* 本文系国家自然科学基金青年项目"面向专利文本中实体关系抽取的远程监督方法研究"(项目编号:71704169)和国家自然科学基金青年项目"大数据挖掘在科技项目查重中的应用"(项目编号:71303223)研究成果之一。
作者简介:曹祺(ORCID:0000-0001-6337-3451),博士后,博士,E-mail:caoqi@istic.ac.cn;赵伟,研究员,博士;张英杰,副研究馆员,博士,;赵树君(ORCID:0000-0001-9310-6135),博士研究生;陈亮(ORCID:0000-0002-3235-9806),助理研究员,博士。
收稿日期:2017-10-16 修回日期:2018-04-05 本文起止页码:74-81 本文责任编辑:杜杏叶

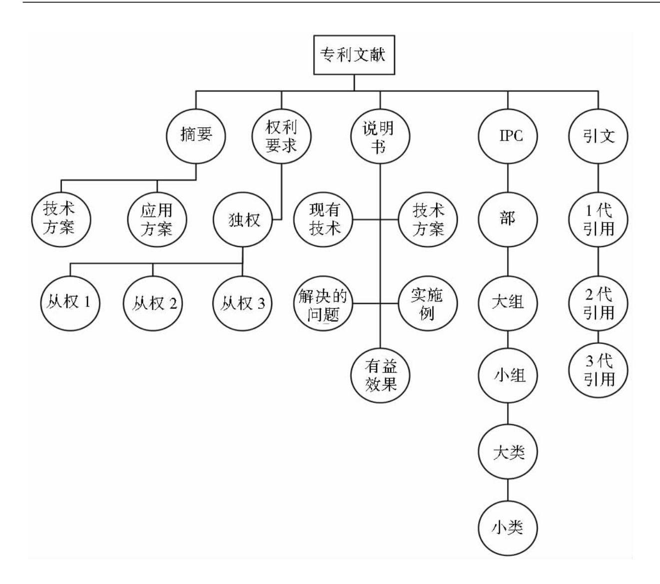



图1　专利结构树示意图[4]

注:本图1参照文献[4]中图1绘制,引用自参考文献[4]

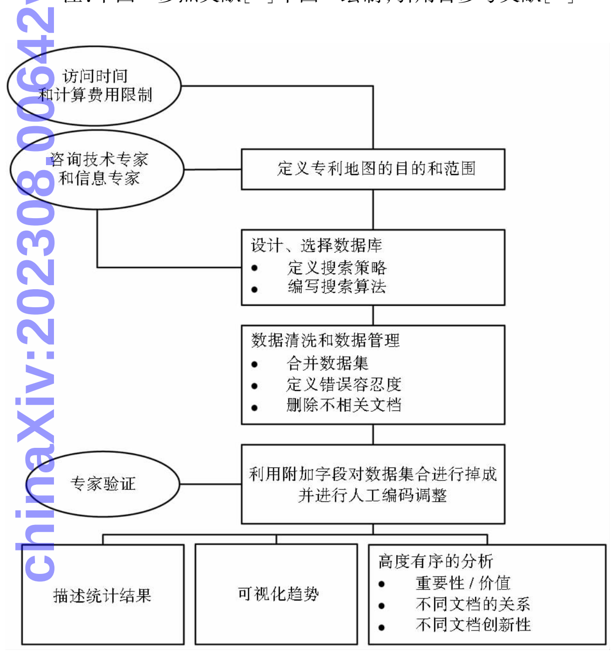


图2　专利相似度分析流程[5]

注:本图2参照文献[5]中图1绘制,引用自参考文献[5]

在数据源算法上主要分为专利数据清洗算法和数据搜索对比算法两类。通过专利自身数据分析主要是找到专利间的各种关系,如研究专利引用网络的引用聚类关系。通过专利和其他业务数据库的跨业务数据库对比是为了方便专利领域专家进行人工分析,见图3。

2.2　专利相似度的研究对象

研究专利相似度分为基于专利关联其他业务数据库进行研究和基于专利文件的自身数据库继续研究。

(1)专利关联其他业务数据库:在这个领域,如 S. Mukherjee 等[1]通过对比论文数据库和专利数据库来研究专利和论文技术一致性和热度问题。J. A. Smith 等[6]通过对比专利发明人的论文分布来对比研究专利质量。李莉等[7]通过对比中文专利和英文专利来研究专利词语的消歧问题。娄岩等[8]通过专利数据和商业数据的对比分析来研究专利的替代性方案。跨库对比研究相似度问题的优势在于原理简单,能得出可论证的结论,但是跨库对比研究的难度是要处理大量的不同的数据库,并且需要有不同行业的相关领域知识和洞察力。

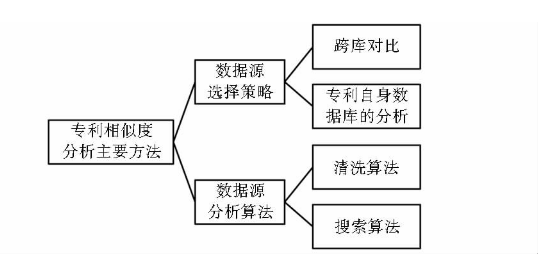


图3　专利相似度分析主要方法

(2)专利自身数据库:基于图1的专利结构树,主要的数据选择如表1[5]所示:

这类研究一般基于指定研究领域的同一类型的文本进行分析,如陈云伟等[9]通过研究专利引用网络,来判断专利发明人之间的合作情况;另一方面也有将专利结构中的不同类型的数据进行分析的,如王鑫等[10]基于分类号和引文的专利相似度测量。朱磊等[11]将专利的文本数据和图像数据进行对比分析,基于形状语义进行外观专利图像的检索。利用专利自身数据库的分析方法更依赖于研究者对专利行业的经验,但是相对于跨业务数据库对比研究而言,处理的数据量会更小更容易。

2.3　基于领域知识的研究方法

常用的相似度检测算法需要基于专利领域知识进行数据清洗,数据清洗的主要目的是除噪,对同义词进行消除歧义,降低数据处理的计算维度,生成对应的实体表示,如王晋等[12]利用最大熵模型对专利文本生成实体。但是数据清洗本身并不是简单的进行无效词表删除,自然语言处理中的一词多义和多词同义本身就很复杂。从应用的场景来说,专利审查员不一定是行业专家,但是对专利规则理解清洗,因此需要专利发明人将专利的语言尽可能的补充相关解释说明以方便专利审查员理解,如陈亮等[13]利用 Knowledge Graph 对专利隐式实体进行补全。在词义歧义消除方面,一般采用传统的自然语言处理方法,如利用 WordNet 相关技术分析词义上下位关系消除歧义或者利用词法、词

表 1 基于专利结构树的分析[5]

通过专利地图的研究问题	专利地图分析类型	分析依据
专利在指定技术、产品、领域的覆盖面	• 技术专利地图 • 技术专利地图的对比分析 • 搜索优先权 • 专利授权通过率分析	专利分类分析/权利要求分析/技术关键词 • 争议专利 • 专利申请书 • 国家和来源分析 • 专利转让人/发明人
专利权如何影响了公司或者机构	• 机构投资组合分析 • 发明人投资租房分析 • 专利性能对比 • 竞争者分析 • 行业分析	专利转让人/发明人 • 争议专利 • 专利申请书 • 国家和来源分析 • 专利分类分析/权利要求类型分析
不同国家或者地域的专利权地图情况	• 区域创新指数 • 创新聚类分析 • 外来技术分析 • 国际专利树分析 • 基于时间的国家地区的专利演化分析	地理位置区域分析 • 专利分类分析/权利要求类型分析 • 争议专利 • 专利申请书 • 国家和来源分析 • 专利转让人/发明人
哪些专利是最重要或者最优价值的专利	• 权利要求结构 • 文献计量分析 • 诉讼分析	• 专利号/权利要求范围 • 前置引用 • 专利家族 • 专利诉讼 • 专利维护费缴纳状况
专利之间的关系	• 文献计量分析 • 网络引用分析 • 语义相似度分析	• 前置引用/后置引用 • 关键词 • 共同发明人 • 专利转让人/发明人 • 专利网络统计分析
专利在创新或者竞争领域产生怎样的影响	• 专利数分析 • 专利权利要求分析 • 专利密度分析 • 统计模型分析	• 权利要求范围 • 争议专利 • 专利分类分析/权利要求类型 • 专利申请书 • 国家和来源分析 • 专利转让人/发明人

注:本表 1 参照文献[5]中表 1 绘制,引用自参考文献[5]

义生成属于词库消除歧义,如姜利雪等[14]利用语义角色生成专利术语词库。另外,由于不同语言本身的词法、语法的差异,需要将专利的句式变简单,方便机器和人工的进一步分析,这类主要采用的方法是 SAO 方法提取,如许海云等[15]在基于 SAO(Subject Action Object)提取算法对专利句式分析进行对比研究。而饶齐等[16]在 SAO 的基础上基于句法分析,结合 SPT (the Shortest Path enclosed Tree)结构进行改进。

数据分析的主要目的是在数据清洗的基础上,利用相关算法进行相似度分析,通过生成专利实体建立专利地图数据库,将新专利和已经建立好的数据库对比判断其相似度和新颖性。目前国内专利审查员主要用的是国家知识产权局的专利检索与服务系统(S 系统),相关检索主要是基于 VSM(Vector Space Model)模型,检索后通过人工判断词语位置和字面相似度来判断技术方案的创新性[17]。该方案用起来原理很简单,但是需要大量的专利审查员进行人工干预。除了 VSM 模型,比较常用的模型还有 LSA(Latent Semantic Analysis)模型,LDA(Latent Dirichlet Allocation)模型。这几类模型对比相似度的原理大致相同,都是将专利采用词袋模型,将专利的每个词进行打分计算词的权重,常用的打分策略是 TF-IDF(Term Frequency-Inverse Document Frequency)算法,然后将一篇专利按照词频和根据 TF-IDF 算法打分建立 VSM 模型,然后对比不同专利文件的在 VSM 模型中的向量夹角来对比相似度,由于 VSM 模型中的向量分布太过稀疏,因此采用 LSA 模型进行 SVD(Singular Value Decomposition)分解降维,如果考虑的主题(Topic)因素,则采用 LDA 模型利用主题进行降维。如陈亮等[18]利用 LDA 模型通过研究专利实体来判断专利演化过程中主题相似度。廖列法等[19]对比了 LDA 模型和 VSM 模型在专利主题相似度分析的正确率和召回率。

尽管 VSM 模型,LSA 模型,LDA 模型应用广泛,相对成熟,但是对结果进行二次分析仍然依赖于专利分

析人员的领域知识,耗费大量人工。

2.4 基于深度学习的研究方法

针对专利相似度分析的专利领域知识需要大量专业人才的问题,本文提出另一种研究思路,即不基于领域知识的研究方法而是基于深度学习的研究方法。

这种新方法的提出主要归功于最近几年神经网络和深度学习相关技术的成熟。神经网络方法和传统的自然语言处理的方法最大的区别是模拟人脑的树触和轴突功能,即信息通过激活函数判断结果后传递学习。而深度学习方法在传统神经网络方法上模拟人脑神经细胞处理信息时的功能无差异性,能够对训练数据进行分片,如广泛应用于图像识别的卷积神经网。同时采用不同的分类器对同一分片的数据进行不同分类的学习,最后将不同分片的学习结果进行合并。在专利相似度分析领域,相关学者引入了神经网络的相关算法进行相似度研究,尤其是无监督学习,武玉英等[20]提出了基于自组织神经网络 SOM(Self Organization Map)进行训练,通过训练生成关键词和专利权重的矩阵进行专利相似度计算和侵权检测。许侃等[21]利用深度学习(Deep Learning)中的 Word2Vec 框架对专利文本进行训练,通过训率判断专利领域词语的相似度。也有采用深度学习相关算法消除歧义,如王琰炎等[22]利用 Word2vec 框架进行词义消歧义。神经网络相关的研究方法相对而言,能减少人工干预,另外训练过程中由于不受传统的自然语言处理中的相关约束,计算效率均有很大提升。

过去常用的自然语言处理的方法更类似人工分类和利用数学模型消除噪音提高精准度,而神经网络的方法,尤其是深度学习的方法,更类似事先不约定业务模型,利用神经网络去生成模型,发现其中规律。尽管此类方法的进一步优化需要相关研究者结合领域知识利用传统的自然语言处理方法进行加工[5]。本文的相关创新也是采用不基于领域知识的研究方法而是基于深度学习的研究方法。本文基于 Doc2Vec 的模型,Doc2Vec 是 Word2Vec 的模型,不过在训练词向量的过程中输入层增加了段落矩阵[23]。Doc2Vec 更适用于处理专利文本因为适用于段落,比如处理专利的摘要。

3 实验设计及结果分析

3.1 实验技术路线

本实验采用的理论基础是假设与假设验证原则,统计法原则和对照与实验对照原则。本文的假设前提是不采用基于专利领域知识进行数据清洗,而采用深度学习的 Doc2Vec 模型,然后结合传统的 TFIDF 模型、LSA 模型和 LDA 模型,对比专利相似度检测结果,找到相近性。具体来说,先通过训练生成各个模型文件和检索文件,然后深度学习模型作为传统模型的实验对照组,利用 TFIDF 模型、LSA 模型和 LDA 模型和 Doc2Vec 模型进行对照。先随机抽出一组专利进行分别对照分析,然后得出假设的结论,再随机进行另一组专利重复实验验证假设,由于实验时验证规律时需要加多的对比,因此统计的时候选取 100 项进行统计,分析其结果并验证规律。

3.2 实验环境及准备

本文主要利用 Gensim 的框架[24]进行实验,实验代码基于 Python 2.7.12,采用的数据库为 MariaDB 数据库,数据库版本为 10.1.21,整个实验代码基于 Gensim 框架 3.0.0,主要的开发环境为 Ubuntu Linux v16,64 位操作系统,处理器为 Intel 的 16 核处理器,运行的内存为 64G。

本研究的语料首先下载了美国专利局(USPTO)2015 年 1 月 1 日至 2017 年 8 月 1 日的专利数据,共计 3 044 956 条专利数据,并且将其导入到数据库。将这些专利的专利号和摘要保留,生成由专利号和对应的专利摘要构成的 CSV 文本语料文件,其数据结构如图 4 所示:

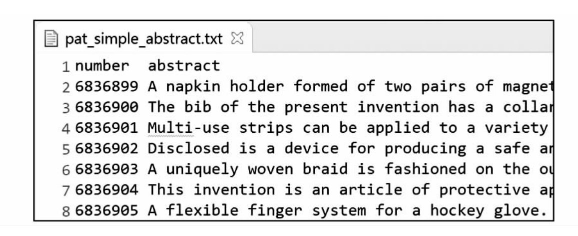


图 4 生成的 CSV 文本语料文件

3.3 实验设计及流程

根据 3.2 节的实验环境,设计了整个实验流程,主要分为训练库生成(Train)、推断测试(Infer)两个部分,训练库生成的流程图如图 5 所示(注:LSA 生成索引的方法叫做 LSI):

当模型生成后,需要利用训练生成的模型与待测试的新的专利进行对比,并得出其专利号,具体的推断测试流程见图 6。

3.4 实验结果分析

根据 3.3 节图 5 和图 6 的实验流程进行训练,对于 TF-IDF 模型 E 组采用 D 文件组的词典长度作为特征数,对于 LDA 模型 F 组和 LSA 模型 G 组的特征数则定义为 10。对于所提出的基于深度学习的 Doc2Vec 方

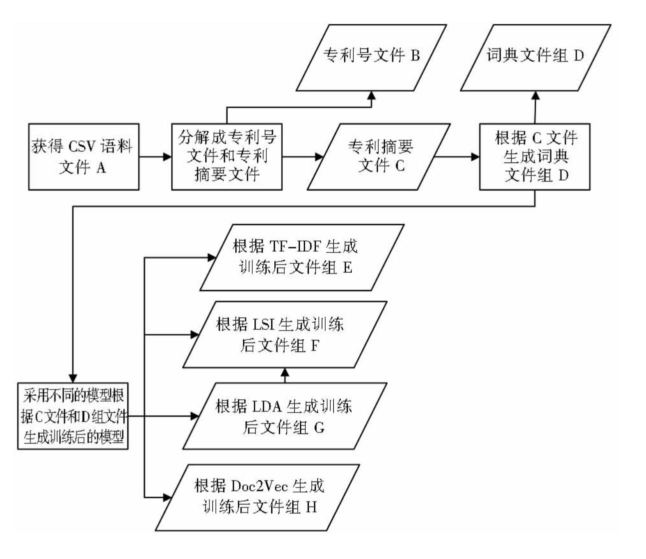


图 5　模型生成流程

法,考虑到性能优化,将 3 044 956 份专利文件进行分块存储,每块参与计算的专利为 327 680 份,对于 Doc2Vec 模型训练时对每个专利文件训练 100 次。具体而言,训练模型的状态如下表 2 所示:

表 2　训练后的文件模型

序号	文件类型	文件大小
1	D 文件组	词库 35.1MB,词典语料库 2G
2	E 文件组	TFIDF 模型文件 25.2MB
3	F 文件组	LSA 模型文件 150MB
4	G 文件组	LDA 模型文件 169MB
5	H 文件组	Doc2Vec 模型文件 12.5GB

为了验证试验的可靠性,本试验随机从 2017 年 9 月 5 日的专利数据库[25]中选择两条专利进行假设及推断,其中一条数据用于假设,另一条数据用于推断测试,结果见表 3。

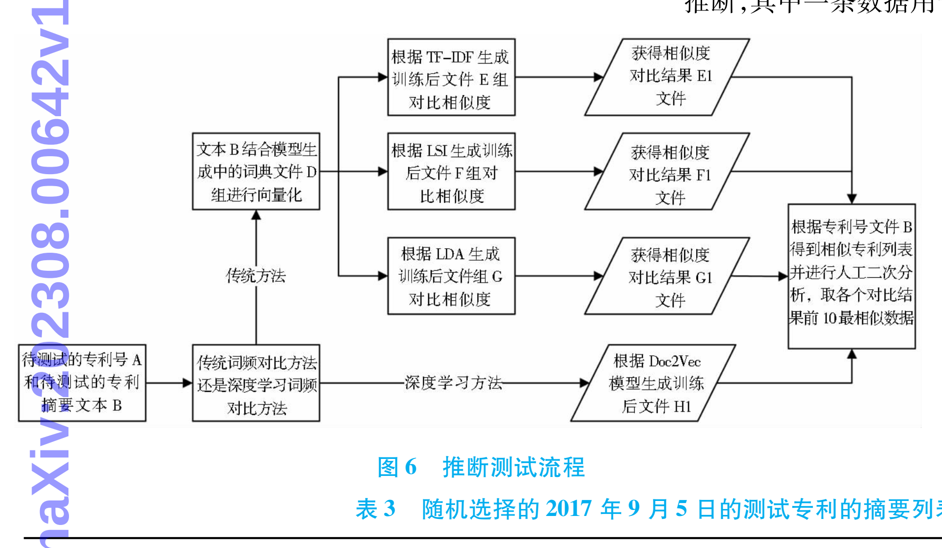


图 6　推断测试流程

随机选择的专利号分别是 9754858、9755578。对比 TF-IDF 生成文件 E1、LSA 生成文件 F1、LDA 模型生成文件 G1、Doc2Vec 模型生成文件 H1,各自的实验结果见表 4。

对比分析前 20 组数据中不同相似度算法得出的共同项情况见表 5。

表 3　随机选择的 2017 年 9 月 5 日的测试专利的摘要列表

专利号	日期	摘要
9754858	2017-09-05	Provided is a gas sensor package, including: a gas sensing element; and a substrate on which the gas sensing element is disposed, in which a through hole corresponding to the gas sensing element is formed.
9755578	2017-09-05	Current-mode control for radio-frequency (RF) power amplifiers. In some embodiments, an RF power amplifier control circuit can include a sensor configured to measure a base current of a power amplifier and generate a sensed current. The control circuit can further include a sensing node configured to receive a reference current and perform a current-mode operation with the sensed current to yield an error current. The control circuit can further include a control loop configured to generate a control signal based on the error current to adjust an operating parameter of the power amplifier.

表 4　对于 9754858 专利文件不同模型的相似度实验结果

排名	E1 中专利号和相似度组 W		F1 中专利号和相似度组 X		G1 中专利号和相似度组 Y		H1 中专利号和相似度组 Z	
1	9314325	0.6425	8853491	0.9900	7151570	0.9999	7874783	0.5317
2	9131564	0.5334	8529514	0.9885	7151742	0.9999	9314325	0.5232
3	7264711	0.5149	8601325	0.9879	7048667	0.9999	7825572	0.4780
4	7780521	0.5017	8355788	0.9873	8011156	0.9998	7405461	0.4686
5	7584294	0.4914	7701684	0.9856	8804567	0.9998	8995191	0.4655
6	7405461	0.4879	9275622	0.9851	7108194	0.9998	9091223	0.4630
7	7825572	0.4879	8969944	0.9846	8347892	0.9998	8226232	0.4626
8	8226232	0.4879	9306556	0.9846	6921204	0.9998	9372451	0.4564
9	D683561	0.4862	9258134	0.9835	D534287	0.9998	8130847	0.4479
10	8233345	0.4846	8988869	0.9834	7923944	0.9998	7999571	0.4477

曹祺，赵伟，张英杰，等.基于 Doc2Vec 的专利文件相似度检测方法的对比研究[J].图书情报工作,2018,62(13):74－81.

(续表 4)

排名	E1 中专利号和相似度组 W		F1 中专利号和相似度组 X		G1 中专利号和相似度组 Y		H1 中专利号和相似度组 Z	
11	9144108	0.4786	8330153	0.9829	9025779	0.9998	9262774	0.4470
12	8619592	0.4734	8548142	0.9828	7517616	0.9997	7542956	0.4436
13	9045346	0.4709	7684629	0.9823	7648748	0.9997	7017519	0.4414
14	7874783	0.4682	7373520	0.9821	7669858	0.9997	8608654	0.4410
15	9202784	0.4668	7437194	0.9819	8243253	0.9997	7684753	0.4405
16	8186819	0.4655	8109960	0.9817	8715534	0.9997	8619592	0.4360
17	8774912	0.4637	7199532	0.9817	7285144	0.9997	8819039	0.4325
18	8799588	0.4619	7449201	0.9809	8026685	0.9997	9177210	0.4315
19	8819039	0.4482	7960537	0.9807	8101977	0.9997	7490473	0.4147
20	8851531	0.4446	8260273	0.9807	7130929	0.9997	8460467	0.4141

表 5　对于 9754858 专利文件不同模型的相似度对比分析

交集对比	W 组	X 组	Y 组	Z 组
W 组		0	0	7
X 组			0	0
Y 组				0
Z 组				

我们发现 W 组(TFIDF 模型)和 Z 组(Doc2Vec 模型)在前 20 项数据中有 7 项相同,但是 W 组和 X 组、Y 组没有共同项。X 组和 Y 组进行 LSA 模型和 LDA 模型也是基于 TF-IDF 模型。

本文的研究发现,如果不进行基于专利领域知识的数据清洗,则 X 组和 Y 组会造成数据没有交集,即没有相似性,但是对于 TF-IDF 模型和 Doc2Vec 模型的相似性效果比 LSA 模型和 LDA 模型较好。

基于这样的假设,我们利用第二份专利(专利号:9755578)进行测试,主要测试是否存在相似性。本文将实验数据增大,由于 X 组和 Y 组都是基于 W 组,但是 Z 组的模型不是基于 X 组,因此进行实验时,取 W 组、X 组和 Y 组前 100 条结果和 Z 组前 20 条数据进行不同模型的相似度对比分析,按照表 4、表 5 的流程对任意两组进行对比,对比分析前 20 组数据中不同相似度算法得出的共同项情况如表 6 所示:

表 6　对于 9755578 专利文件不同模型的相似度对比分析

交集对比	W 组	X 组	Y 组	Z 组
W 组		1	1	3
X 组			0	0
Y 组				0
Z 组				

表 5、表 6 的实验结果进一步验证了本文根据表 4 实验数据推断出的假设,即对专利相似度进行比较的时候,TF-IDF 模型和 Doc2Vec 模型的相似性检测要优于 LSA 模型和 LDA 模型,实验结果表明,如果不做基于专利领域知识的数据清洗工作,基于深度学习的 Doc2Vec 方法所得出的结果和 TF-IDF 相近,目前行业采用的相似度检测方法主要也是基于专利领域知识选择好数据采用 TF-IDF 方法检测。

4　结论

4.1　研究价值

本文提出了基于深度学习的 Doc2Vec 专利相似度分析方法,并采用 Doc2Vec 进行了案例分析,并将结果与传统的 TF-IDF 模型、LSA 模型、LDA 模型相比较,开展了假设与假设验证实验,和对照实验。本文提出的创新点是用深度学习相关的模型和算法来对比研究专利相似度问题,基于深度学习的 Doc2Vec 专利相似度分析方法在于不需要研究人员有较多的专利领域知识,过去传统的研究思路是将研究对象进行分类,但是由于研究对象数据巨大,实体定义和数据清洗需要耗费大量的工作时间,同时实体定义依赖于专家系统,需要大量的具备专利领域知识的专家。而本文所提出的采用深度网络的 Doc2Vec 新方法不需要基于专利领域知识进行数据清洗而得到了与传统方法相类似的结果。

另外本文并未做太多的数据定义和数据清洗,采用了较长时间的机器训练,这样的目的也是防止信息丢失。在训练 TF-IDF 模型、LSA 模型,LDA 模型和 Doc2Vec 模型时,本文的测试专利文件进行打分时会和 327 680 份专利文件进行对比,也为相关研究提供了另一种研究思路,尽管提高无干预的机器训练的计算强度和计算时间,但是尽可能采用无监督学习的模式,而不是过多的行业专家进行数据库前期加工,尽可能的纯计算机自动操作,减少人工工作量。

4.1　应用价值

在应用上,本文可以为专利地图的辅助生成提供

基于 Doc2Vec 的模型,传统专利地图的生成主要是基于 TF-IDF 模型和 LSA 模型,然后不断地利用大量具备专利领域知识的研究人员进行词库划分(如专利词义划分),而本文提供的思路是让不具备专利领域的研究人员进行专利分析,专利的词义划分不是基于人工而是基于深度学习生成的模型,尽管此类生成的模型可能结果有歧义,不如纯人工方法检测的结果清晰,但是可以极大地节约专利代理人和专利分析人员的精力。并且从专利侵权领域来说,由于 Doc2Vec 模型用于查询的时候会通过训练学习语义,因此在专利侵权领域能发现 TFIDF 和 LSA 模型发现不了的侵权案例,方便相关企业在竞争格局和侵权分析时,对侵权专利的结果进行补充分析,提升服务效率。

参考文献:

[1] MUKHERJEE S, ROMERO D M, JONES B, et al. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: the hotspot [J]. Science advances, 2017, 3(4): e1601315.
[2] PADMANABHAN S, AMIN T, SAMPAT B, et al. Intellectual property, technology transfer and manufacture of low-cost HPV vaccines in India [J]. Nature biotechnology, 2010, 28(7): 671-678.
[3] 王曰芬, 谢寿峰, 邱玉婷. 面向预警的专利文献相似度研究的意义及现状 [J]. 情报理论与实践, 2014, 37(7): 135-140.
[4] 王秀红,袁艳,赵志程,等. 专利文献的结构树模型及其在相似度计算中的应用 [J]. 情报理论与实践, 2015, 38(3): 107-111.
[5] BUBELA T, GOLD E R, GRAFF G D, et al. Patent landscaping for life sciences innovation: toward consistent and transparent practices [J]. Nature biotechnology, 2013, 31(3): 202-206.
[6] SMITH J A, ARSHAD Z, THOMAS H, et al. Evidence of insufficient quality of reporting in patent landscapes in the life sciences [J]. Nature biotechnology, 2017, 35(3): 210-214.
[7] 李莉, 刘知远, 孙茂松. 基于中英平行专利语料的短语复述自动抽取研究 [J]. 中文信息学报, 2013, 27(6): 151-158.
[8] 娄岩, 张赏, 黄鲁成. 基于专利分析的替代性技术选择研究 [J]. 科技管理研究, 2015, 35(20): 150-154.
[9] 陈云伟, 方曙. 专利权人关联网络的社会网络分析方法研究 [J]. 图书情报知识, 2011(3):58-66.
[10] 王鑫, 赵蕴华, 高芳. 基于分类号和引文的专利相似度测量方法研究 [J]. 数字图书馆论坛, 2015(01):57-62.
[11] 朱磊, 金海, 郑然, 等. 基于形状语义的外观设计专利检索 [J]. 计算机辅助设计与图形学学报, 2013, 25(3): 372-380.
[12] 王晋, 孙涌, 王璁玮. 基于领域本体的文本相似度算法 [J]. 苏州大学学报: 工科版, 2011, 31(3): 13-17.
[13] 陈亮, 张海超, 杨冠灿, 等 利用 Knowledge Graph 的专利表示方法及其应用 [J]. 图书情报工作, 2017, 61(9): 123-129.
[14] 姜利雪, 季铎, 蔡东风. 专利中基于语义角色的术语相似度计算方法 [J]. 中文信息学报, 2016, 30(4): 37-43.
[15] 许海云, 王振蒙, 胡正银, 等. 利用专利文本分析识别技术主题的关键技术研究综述 [J]. 情报理论与实践, 2016,39(11): 131-137.
[16] 饶齐, 王裴岩, 张桂平. 面向中文专利 SAO 结构抽取的文本特征比较研究 [J]. 北京大学学报 (自然科学版), 2015, 51(2): 349-356.
[17] 杨宏章, 付静. 利用专利文本结构化特征构建专利信息智能语义检索系统的方法 [J]. 情报理论与实践, 2015, 38(4): 136-138.
[18] 陈亮, 杨冠灿, 张静, 等. 面向技术演化分析的多主路径方法研究 [J]. 图书情报工作, 2015, 59(10): 124-130,115.
[19] 廖列法, 勒孚刚, 朱亚兰. LDA 模型在专利文本分类中的应用 [J]. 现代情报, 2017, 37(3): 35-39.
[20] 武玉英, 马羽翔, 翟东升. 基于 SOM 的中文专利侵权检测研究 [J]. 情报杂志, 2014, 33(2): 33-39.
[21] 许侃,林原,曲忱,等. 专利查询扩展的词向量方法研究 [J]. 计算机科学与探索, 2017(9):1-9.
[22] 王琰炎, 王裴岩. 一种用于专利实体的实体消歧方法 [J]. 沈阳航空航天大学学报, 2015, 32(1): 77-83.
[23] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of the 31st international conference on machine learning (ICML-14). [EB/OL]. [2014-01-27] http://proceedings.mlr.press/v32/le14.html.
[24] ŘEHŮŘEK R, PETR S. Software framework for topic modelling with large corpora. [EB/OL]. [2018-05-22]. https://is.muni.cz/publication/884893/en%7D%7D%2C%20language=%7BEnglish? lang=en.
[25] UNITED STATES PATENT TREADEMARK OFFICE. Patent Grant Full Text Data/XML Version 4.5 ICE (JAN 2017 - DEC 2017) [EB/OL]. [2017-12-26]. https://bulkdata.uspto.gov/data/patent/grant/redbook/fulltext/2017/.

作者贡献说明:

曹祺:负责论文思路及框架构建、主体内容撰写、实验的开展和结果分析与讨论;赵伟:负责文献调研、参与论文框架的构建和实验结果的分析与讨论;
张英杰:参与模型的预处理、分析与讨论;
赵树君:参与实验模型的数据采集、预处理以及结果分析、结果的完善;
陈亮:负责数据校对,并在论文撰写过程中提出修改意见。

曹祺，赵伟，张英杰，等. 基于 Doc2Vec 的专利文件相似度检测方法的对比研究[J]. 图书情报工作,2018,62(13):74－81.

Comparative Study of Patent Documents Similarity Detection on Deep Learning of Doc2Vec Based Methods

Cao Qi[1] Zhao Wei[1] Zhang Yingjie[2] Zhao Shujun[3] Chen Liang[4]

[1,2,4] Institute of Scientific and Technical Information of China, Beijing 100038

[3] Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] Patent similarity detection assists the formulation of the national innovation strategy planning macroscopically, finds hotspots in China and all over the world, and deals with patent rogues in other countries. Microscopically, patent similarity detection provides support for patent inventors, patent examiners and patentees. [**Method/process**] A new method was proposed based on deep learning of Doc2Vec model, with patent corpus based on no data clearance of domain knowledge. Then typical patents were randomly selected to carry on similarity detection by this new method, and the results with traditional similarity detection models were compared. [**Result/conclusion**] According to experimental results, the new deep learning of Doc2Vec method and TFIDF model has similary results which both of the model's patent corpus all based on no data clearance of domain knowledge. The new method requires less professional skill in specific domain knowledge, and didn't require the process of data clearance. It can givesa new intelligent support tool for patent infringement and patent investigation, reduce the research threshold and workload, and improve service efficiency.

Keywords: patent similarity deep-learning Doc2Vec

关于在学术论文署名中常见问题或错误的诚信提醒

恪守科研道德是从事科技工作的基本准则，是履行党和人民所赋予的科技创新使命的基本要求。中国科学院科研道德委员会办公室根据日常科研不端行为举报中发现的突出问题，总结当前学术论文署名中的常见问题和错误，予以提醒，倡导在科研实践中的诚实守信行为，努力营造良好的科研生态。

提醒一：论文署名不完整或者夹带署名。应遵循学术惯例和期刊要求，坚持对参与科研实践过程并做出实质性贡献的学者进行署名，反对进行荣誉性、馈赠性和利益交换性署名。

提醒二：论文署名排序不当。按照学术发表惯例或期刊要求，体现作者对论文贡献程度，由论文作者共同确定署名顺序。反对在同行评议后、论文发表前，任意修改署名顺序。部分学科领域不采取以贡献度确定署名排序的，从其规定。

提醒三：第一作者或通讯作者数量过多。应依据作者的实质性贡献进行署名，避免第一作者或通讯作者数量过多，在同行中产生歧义。

提醒四：冒用作者署名。在学者不知情的情况下，冒用其姓名作为署名作者。论文发表前应让每一位作者知情同意，每一位作者应对论文发表具有知情权，并认可论文的基本学术观点。

提醒五：未利用标注等手段，声明应该公开的相关利益冲突问题。应根据国际惯例和相关标准，提供利益冲突的公开声明。如资金资助来源和研究内容是否存在利益关联等。

提醒六：未充分使用志（致）谢方式表现其他参与科研工作人员的贡献，造成知识产权纠纷和科研道德纠纷。

提醒七：未正确署名所属机构。作者机构的署名应为论文工作主要完成机构的名称，反对因作者所属机构变化，而不恰当地使用变更后的机构名称。

提醒八：作者不使用其所属单位的联系方式作为自己的联系方式。不建议使用公众邮箱等社会通讯方式作为作者的联系方式。

提醒九：未引用重要文献。作者应全面系统了解本科研工作的前人工作基础和直接相关的重要文献，并确信对本领域代表性文献没有遗漏。

提醒十：在论文发表后，如果发现文章的缺陷或相关研究过程中有违背科研规范的行为，作者应主动声明更正或要求撤回稿件。

院属各单位应根据以上提醒，结合本单位学科特点和学术惯例，对科研人员进行必要的教育培训，让每一位科研工作者对学术论文署名保持高度的责任心，珍惜学术荣誉、抵制学术不端行为，将科研诚信贯穿于学术生涯始终。

来源：中国科学院监督与审计局